МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Экономический факультет

УТВЕРЖДАЮ:
Проректор по учебной работе,
качеству образования – первый
проректор
_____ Т.А. Хагуров
подпись
«26» мая 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.17 Анализ Big Data

(код и наименование дисциплины в соответствии с учебным планом)

Направление подготовки: 27.03.03 Системный анализ и управление

(код и наименование направления подготовки/специальности)

Направленность (профиль):
Интеллектуальная бизнес-аналитика и управление экономическими процессами

(наименование направленности (профиля) / специализации)

Форма обучения: _____очная_____

(очная, очно-заочная, заочная)

Квалификация: бакалавр

Краснодар 2023

Рабочая программа дисциплины Б1.В.17 Анализ Big Data составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки / специальности 27.03.03 Системный анализ и управление

Программу составил(и): И.В. Ариничев, к.э.н., доцент _

Рабочая программа дисциплины Б1.В.05 «Интеллектуальный анализ данных» утверждена на заседании кафедры «экономики и управления инновационными системами» протокол № 5 «18» апреля 2023г.

Заведующий кафедрой экономики и
управления инновационными системами Литвинский К. О. _____

Утверждена на заседании учебно-методической комиссии экономического факультета протокол № 8 «19» мая 2023 г.
Председатель УМК факультета/института Дробышевская Л.Н. ____

Рецензенты:
Шевченко И.В., д-р экон. наук, профессор, зав. каф. мировой экономики и менеджмента, декан экономического факультета ФГБОУ ВО «КубГУ»

Ксенофонтов В.И., д.э.н., профессор, директор Краснодарского ЦНТИфилиала ФГБУ РЭА Минэнерго РФ

1 Цели и задачи изучения дисциплины (модуля)

1.1 Цель освоения дисциплины

Цель освоения дисциплины Анализ Big Data состоит в формировании знаний, умений и навыков (компетенций) по одному из приоритетных в современных информационных технологиях направлению - аналитической обработке больших данных.

1.2 Задачи дисциплины

1. ознакомление бакалавров с основными принципами машинного обучения - а именно, видами задач машинного обучения, классами моделей (линейные, логические, нейросетевые), метриками качествами и подходами к предварительной обработке данных;
2. формирование у бакалавров практических навыков сбора, обработки данных и решения социально-экономических задач анализа данных на языке Python;
3. формирование у бакаларов представления о технических и методологических средствах анализа больших данных, обеспечивающих хранение и управление объемом данных в сотни терабайт или петабайт, которые обычные РБД не позволяют эффективно использовать;

1.3 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина Б1.В.17 «Анализ Big Data» относится к части, формируемой участниками образовательных отношений Блока 1 "Дисциплины (модули)" учебного плана. В соответствии с рабочим учебным планом дисциплина изучается на 3 курсе по форме обучения. Вид промежуточной аттестации: экзамен.

Перечень предшествующих дисциплин, необходимых для ее изучения:

- Линейная алгебра и аналитическая геометрия;
- Дискретная математика и математическая логика;
- Data Mining
- Программирование на языке Python

Перечень последующих дисциплин, для которых данная дисциплина является предшествующей в соответствии с учебным планом:

- Технологическая (проектно-технологическая) практика.

1.4 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих компетенций:

Код и наименование индикатора* достижения компетенции	Результаты обучения по дисциплине
ПК-2 Способен анализировать и исследовать большие данные с использованием существующей в организации методологической и технологической инфраструктуры	
ИПК-2.9. Анализирует большие данные с использованием современных методов и имеющейся технолого-методологической инфраструктуры	*Знает*: типы анализа больших данных, виды аналитики, современные методы и инструментальные средства анализа больших данных; *Знает*: возможности использования свободно распространяемого программного обеспечения для анализа больших данных; *Знает*: современный опыт использования, теоретические и прикладные основы анализа больших данных.
	Умеет: проводить сравнительный анализ методов и инструментальных средств анализа больших данных; *Умеет*: планировать аналитические работы с использованием технологий больших данных; *Умеет*: проводить анализ больших данных, осуществлять их интеграцию и преобразование
	Трудовое действие: выбор методов и инструментальных средств анализа больших данных для проведения аналитических работ; *Трудовое действие*:

Код и наименование индикатора* достижения компетенции	Результаты обучения по дисциплине
	Реализация интеллектуальных алгоритмов на структурированных и неструктурированных данных

Результаты обучения по дисциплине достигаются в рамках осуществления всех видов контактной и самостоятельной работы обучающихся в соответствии с утвержденным учебным планом.

Индикаторы достижения компетенций считаются сформированными при достижении соответствующих им результатов обучения.

2. Структура и содержание дисциплины

2.1 Распределение трудоёмкости дисциплины по видам работ

Общая трудоёмкость дисциплины составляет 4 зачетные единиц (144 часа), их распределение по видам работ представлено в таблице

Виды работ		Всего часов	Форма обучения			
			очная		очно-заочная	заочная
			7 семестр (часы)	7 семестр (часы)	X семестр (часы)	X курс (часы)
Контактная работа, в том числе:			**56,3**	**56,3**		
Аудиторные занятия (всего):			**50**	**50**		
занятия лекционного типа			34	34		
лабораторные занятия			16	16		
практические занятия						
семинарские занятия						
Иная контактная работа:			**6,3**	**6,3**		
Контроль самостоятельной работы (КСР)			6	6		
Промежуточная аттестация (ИКР)			0,3	0,3		
Самостоятельная работа, в том числе:			**52**	**52**		
Самостоятельное изучение разделов, самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий, подготовка к лабораторным и практическим занятиям, коллоквиумам и т.д.)			52	52		
Контроль:			**35,7**	**35,7**		
Подготовка к экзамену			35,7	35,7		
Общая трудоемкость	**час.**		**144**	**144**		
	в том числе контактная работа		**56,3**	**56,3**		
	зач. ед		**4**	**4**		

2.2 Содержание дисциплины

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.

Разделы (темы) дисциплины, изучаемые в 7 семестре *(**очная** форма обучения)*

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1.	Big Data (большие данные): современные подходы к обработке и хранению	14	4		2	8
2.	Программное обеспечение в области анализа больших данных.	14	4		2	8
3.	Способы получения данных из сети Интернет	14	4		2	8
4.	Введение в машинное обучение	14	4		2	8
5.	Задача классификации. Метрические методы. Логические методы.	16	6		2	8
6.	Линейные модели. Введение в нейронные сети.	14	6		2	6
7.	Обучение без учителя	16	6		4	6
	ИТОГО по разделам дисциплины	102	34		16	52
	Контроль самостоятельной работы (КСР)	6				
	Промежуточная аттестация (ИКР)	0,3				
	Подготовка к текущему контролю	35,7				
	Общая трудоемкость по дисциплине	**144**				

Примечание: Л – лекции, ПЗ – практические занятия / семинары, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

2.3 Содержание разделов (тем) дисциплины

2.3.1 Занятия лекционного типа

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1.	Big Data (большие данные): современные подходы к обработке и хранению	Общие понятия и признаки Big Data. Подходы к обработке и хранению Big Data. Извлечение и визуализация данных. Этапы моделирования. Процесс построения моделей. Формы представления данных, виды и типы данных.	*Т, О*
2.	Программное обеспечение в области анализа больших данных.	Полнотекстовый поиск. Параллельные запросы. Технологии поиска и интеграции. Программные средства.	*Т, О*
3.	Способы получения данных из сети Интернет	Основы работы браузера и протокола HTTP. Формы запросов -get, post. Введение в парсинг сайтов: запросы и ответы, библиотека requests, Selenium. Основы структуры HTML. Библиотека BeautifulSoap, работа с API на примере VK API	*Т, О*
4.	Введение в машинное обучение	Введение. Постановки задач в машинном обучении. Обучение с учителем и без. Классификация, регрессия, ранжирование, кластеризация. Обучающая и тестовая выборки. Проблема переобучения. Кросс-валидация.	*Т, О*
5.	Задача классификации. Метрические методы. Логические методы.	Метод ближайших соседей. Решающие деревья. Бэггинг и метод случайных подпространств. Случайные леса. Бустинг. Градиентный бустинг над решающими деревьями. Модель xgboost.	*Т, О*
6.	Линейные модели. Введение в нейронные сети.	Персептрон. Метод опорных векторов. Задача оценивания вероятностей, логистическая регрессия. Структура нейронной сети. Обратное распространение ошибки. Применение нейросетей для анализа изображений: свёрточные слои.	*Т, О*
7.	Обучение без учителя	Кластеризация данных, задачи обобщения, обнаружения аномалий, сокращения размерности, визуализации данных.	*Т, О*

2.3.2 Занятия семинарского типа (практические / семинарские занятия/ лабораторные работы)

№	Наименование раздела (темы)	Тематика занятий/работ	Форма текущего контроля
1.	Big Data (большие данные): современные подходы к обработке и хранению	Поиск и определение Big Data Хранение больших данных	*ЛР*
2.	Программное обеспечение в области анализа больших данных.	Аналитические платформы: классификация и особенности применения.	*ЛР*
3.	Способы получения данных из сети Интернет	Парсинг сайтов. Работа с API.	*ЛР*
4.	Введение в машинное обучение	Первичный анализ данных с Pandas. Визуальный анализ данных с Python.	*ЛР*
5.	Задача классификации. Метрические методы. Логические методы.	Метод ближайших соседей. Логистическая регрессия. Композиции: бэггинг, случайный лес.	*ЛР*
6.	Линейные модели. Введение в нейронные сети.	Построение и отбор признаков. Приложения в задачах обработки текста, изображений и геоданных.	*ЛР*
7.	Обучение без учителя	Обучение без учителя: PCA, кластеризация.	*ЛР*

Защита лабораторной работы (ЛР), выполнение курсового проекта (КП), курсовой работы (КР), расчетно-графического задания (РГЗ), написание реферата (Р), эссе (Э), коллоквиум (К), тестирование (Т) и т.д.

2.3.3 Примерная тематика курсовых работ (проектов)

Не предусмотрено

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	Занятия лекционного и семинарского типа	Методические указания для подготовки к занятиям лекционного и семинарского типа. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
2	Выполнение самостоятельной работы обучающихся	Методические указания по выполнению самостоятельной работы обучающихся. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
3	Выполнение расчетно-графических заданий	Методические указания по выполнению расчетно-графических заданий. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
4	Выполнение лабораторных работ	Методические указания по выполнению лабораторных работ. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года.. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya
10	Интерактивные методы обучения	Методические указания по интерактивным методам обучения. Утверждены на заседании Совета экономического факультета ФГБОУ ВО «КубГУ». Протокол № 1 от 30 августа 2018 года. Режим доступа: https://www.kubsu.ru/ru/econ/metodicheskie-ukazaniya

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,
– в форме электронного документа,
– в форме аудиофайла,
– в печатной форме на языке Брайля.
Для лиц с нарушениями слуха:
– в печатной форме,
– в форме электронного документа.
Для лиц с нарушениями опорно-двигательного аппарата:
– в печатной форме,
– в форме электронного документа,
– в форме аудиофайла.
Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. Образовательные технологии, применяемые при освоении дисциплины (модуля)

В ходе изучения дисциплины предусмотрено использование следующих образовательных технологий: лекции, практические занятия, подготовка письменных аналитических работ, самостоятельная работа студентов.

Компетентностный подход в рамках преподавания дисциплины реализуется в использовании интерактивных технологий и активных методов (проектных методик, мозгового штурма, разбора конкретных ситуаций) в сочетании с внеаудиторной работой.

Информационные технологии, применяемые при изучении дисциплины: использование информационных ресурсов, доступных в информационно-телекоммуникационной сети Интернет.

Адаптивные образовательные технологии, применяемые при изучении дисциплины – для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

4. Оценочные средства для текущего контроля успеваемости и промежуточной аттестации

Оценочные средства предназначены для контроля и оценки образовательных достижений обучающихся, освоивших программу учебной дисциплины «*Анализ Big Data*».

Оценочные средства включает контрольные материалы для проведения **текущего контроля** в форме *тестовых заданий, расчетно-графических заданий, контрольных работ* и **промежуточной аттестации** в форме вопросов и заданий к экзамену.

Структура оценочных средств для текущей и промежуточной аттестации

№ п/п	Код и наименование индикатора (в соответствии с п. 1.4)	Результаты обучения (в соответствии с п. 1.4)	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
1	ИПК-2.9. Анализирует большие данные с использованием современных методов и имеющейся технолого-методологической инфраструктуры	*Знает*: типы анализа больших данных, виды аналитики, современные методы и инструментальные средства анализа больших данных; *Знает*: возможности использования свободно распространяемого программного обеспечения для анализа больших данных;	*Тест, защита лабораторных рабо*	Вопрос на экзамене 1-20

		Знает: современный опыт использования, теоретические и прикладные основы анализа больших данных. *Умеет*: проводить сравнительный анализ методов и инструментальных средств анализа больших данных; *Умеет*: планировать аналитические работы с использованием технологий больших данных; *Умеет*: проводить анализ больших данных, осуществлять интеграцию и преобразование данных в ходе работ по анализу больших данных Трудовое действие: выбор методов и инструментальных средств анализа больших данных для проведения аналитических работ; Трудовое действие: Мониторинг эффективности работы аналитики больших данных.		

Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Примерный перечень вопросов и заданий

Тест на тему «Загрузка и знакомство с данными»

Для работы вам понадобятся предобработанные данные на kaggle: «Прогноз популярности статьи на Хабре». Скачайте данные https://drive.google.com/file/d/1nV2qV9otN3LnVSDqy95hvpJdb6aWtATk/view, загрузите первые 1500 наблюдений в Pandas и ответьте на следующие вопросы:

1. В каком месяце (и какого года) было больше всего публикаций?
2. март 2016
3. март 2015
4. апрель 2015
5. апрель 2016

2. Проанализируйте публикации в месяце из предыдущего вопроса
Выберите один или несколько вариантов:

1. Один или несколько дней сильно выделяются из общей картины
2. На хабре всегда больше статей, чем на гиктаймсе

3. По субботам на гиктаймс и на хабрахабр публикуют примерно одинаковое число статей

4. Подсказки: постройте график зависимости числа публикаций от дня; используйте параметр hue; не заморачивайтесь сильно с ответами и не ищите скрытого смысла :)

3. Когда лучше всего публиковать статью?

1. Больше всего просмотров набирают статьи, опубликованные в 12 часов дня
2. У опубликованных в 10 утра постов больше всего комментариев
3. Больше всего просмотров набирают статьи, опубликованные в 6 часов утра
4. Максимальное число комментариев на гиктаймсе набрала статья, опубликованная в 9 часов вечера
5. На хабре дневные статьи комментируют чаще, чем вечерние

4. Кого из топ-20 авторов чаще всего минусуют?

1. @Mordatyj
2. @Mithgol
3. @alizar
4. @ilya42

5. Сравните субботы и понедельники¶

Правда ли, что по субботам авторы пишут в основном днём, а по понедельникам — в основном вечером?

1. Правда
2. Неправда

Для продолжения скачайте данные из репозитория UCI **Adult** и ответьте на следующие вопросы

6. Сколько мужчин и женщин (признак *sex*) представлено в этом наборе данных?

 # Ваш код и ответ здесь

7. Каков средний возраст (признак *age*) женщин?

 # Ваш код и ответ здесь

8. Какова доля граждан Германии (признак *native-country*)?

 # Ваш код и ответ здесь

9. Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50K в год (признак *salary*) и тех, кто получает менее 50K в год?

 # Ваш код и ответ здесь

10. Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак *education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters* или *Doctorate*)

 # Ваш код и ответ здесь

11. Выведите статистику возраста для каждой расы (признак *race*) и каждого пола. Используйте *groupby* и *describe*. Найдите таким образом максимальный возраст мужчин расы *Amer-Indian-Eskimo*.

 # Ваш код и ответ здесь

12. Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак *marital-status*)? Женатыми считаем тех, у кого *marital-status* начинается с *Married* (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.

 # Ваш код и ответ здесь

13. Какое максимальное число часов человек работает в неделю (признак *hours-per-week*)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?

 # Ваш код и ответ здесь

14. Посчитайте среднее время работы (*hours-per-week*) зарабатывающих мало и много (*salary*) для каждой страны (*native-country*).

 # Ваш код и ответ здесь

Лабораторная работа на тему: «Метод главных компонент (PCA), кластеризация»

Мы будем работать с набором данных Samsung Human Activity Recognition. Скачайте данные отсюда. Данные поступают с акселерометров и гироскопов мобильных телефонов Samsung Galaxy S3 (подробнее про признаки – по ссылке на UCI выше), также известен вид активности человека с телефоном в кармане – ходил ли он, стоял, лежал, сидел или шел вверх/вниз по лестнице.

Вначале мы представим, что вид активности нам неизвестнен, и попробуем кластеризовать людей просто на основе имеющихся признаков. Затем решим задачу определения вида физической активности именно как задачу классификации.

Заполните код в клетках (где написано "Ваш код здесь") и ответьте на вопросы.

Подготовительный этап (загрузка библиотек и стилей)

```
import numpy as np
import pandas as pd
import seaborn as sns
from tqdm import tqdm_notebook

%matplotlib inline
from matplotlib import pyplot as plt

plt.style.use(['seaborn-darkgrid'])
plt.rcParams['figure.figsize'] = (12, 9)
plt.rcParams['font.family'] = 'DejaVu Sans'

from sklearn import metrics
from sklearn.cluster import AgglomerativeClustering, KMeans, SpectralClusteri
ng
from sklearn.decomposition import PCA
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.svm import LinearSVC

RANDOM_STATE = 17
X_train = np.loadtxt("../../data/samsung_HAR/samsung_train.txt")
```

```
y_train = np.loadtxt("../../data/samsung_HAR/samsung_train_labels.txt").astyp
e(int)

X_test = np.loadtxt("../../data/samsung_HAR/samsung_test.txt")
y_test = np.loadtxt("../../data/samsung_HAR/samsung_test_labels.txt").astype(
int)
# Проверим размерности
assert(X_train.shape == (7352, 561) and y_train.shape == (7352,))
assert(X_test.shape == (2947, 561) and y_test.shape == (2947,))
```

Для кластеризации нам не нужен вектор ответов, поэтому будем работать с объединением обучающей и тестовой выборок. Объедините *X_train* с *X_test*, а *y_train* – с *y_test*.

```
# Ваш код здесь
X =
y =
```

Определим число уникальных значений меток целевого класса.

```
np.unique(y)
array([1, 2, 3, 4, 5, 6])
n_classes = np.unique(y).size
```

Отмасштабируйте выборку с помощью StandardScaler с параметрами по умолчанию.

```
# Ваш код здесь
scaler =
X_scaled =
```

Понижаем размерность с помощью PCA, оставляя столько компонент, сколько нужно для того, чтобы объяснить как минимум 90% дисперсии исходных (отмасштабированных) данных. Используйте отмасштабированную выборку и зафиксируйте random_state (константа RANDOM_STATE).

```
# Ваш код здесь
pca =
X_pca =
```

Вопрос 1:
Какое минимальное число главных компонент нужно выделить, чтобы объяснить 90% дисперсии исходных (отмасштабированных) данных?

```
# Ваш код здесь
```

Варианты:

- 56
- 65
- 66
- 193

Вопрос 2:
Сколько процентов дисперсии приходится на первую главную компоненту? Округлите до целых процентов.

Варианты:

- 45
- 51
- 56
- 61

```
# Ваш код здесь
```

Визуализируйте данные в проекции на первые две главные компоненты.

```
# Ваш код здесь
plt.scatter(, , c=y, s=20, cmap='viridis');
```

Вопрос 3:
Если все получилось правильно, Вы увидите сколько-то кластеров, почти идеально отделенных друг от друга. Какие виды активности входят в эти кластеры?

Ответ:

- 1 кластер: все 6 активностей
- 2 кластера: (ходьба, подъем вверх по лестнице, спуск по лестнице) и (сидение, стояние, лежание)
- 3 кластера: (ходьба), (подъем вверх по лестнице, спуск по лестнице) и (сидение, стояние, лежание)
- 6 кластеров

Сделайте кластеризацию данных методом KMeans, обучив модель на данных со сниженной за счет PCA размерностью. В данном случае мы подскажем, что нужно искать именно 6 кластеров, но в общем случае мы не будем знать, сколько кластеров надо искать.

Параметры:

- **n_clusters** = n_classes (число уникальных меток целевого класса)
- **n_init** = 100
- **random_state** = RANDOM_STATE (для воспроизводимости результата)

Остальные параметры со значениями по умолчанию.

```
# Ваш код здесь
```

Визуализируйте данные в проекции на первые две главные компоненты. Раскрасьте точки в соответствии с полученными метками кластеров.

```
# Ваш код здесь
plt.scatter(, , c=cluster_labels, s=20,  cmap='viridis');
```

Посмотрите на соответствие между метками кластеров и исходными метками классов и на то, какие виды активностей алгоритм KMeans путает.

```
tab = pd.crosstab(y, cluster_labels, margins=True)
tab.index = ['ходьба', 'подъем вверх по лестнице',
             'спуск по лестнице', 'сидение', 'стояние', 'лежание', 'все']
tab.columns = ['cluster' + str(i + 1) for i in range(6)] + ['все']
tab
```

Видим, что каждому классу (т.е. каждой активности) соответствуют несколько кластеров. Давайте посмотрим на максимальную долю объектов в классе, отнесенных к какому-то одному кластеру. Это будет простой метрикой, характеризующей, насколько легко класс отделяется от других при кластеризации.

Пример: если для класса "спуск по лестнице", в котором 1406 объектов, распределение кластеров такое:

- кластер 1 – 900
- кластер 3 – 500
- кластер 6 – 6,

то такая доля будет 900 / 1406 ≈≈ 0.64.

Вопрос 4:
Какой вид активности отделился от остальных лучше всего в терминах простой метрики, описанной выше?

Ответ:

- ходьба
- стояние
- спуск по лестнице
- перечисленные варианты не подходят

Видно, что kMeans не очень хорошо отличает только активности друг от друга. Используйте метод локтя, чтобы выбрать оптимальное количество кластеров. Параметры алгоритма и данные используем те же, что раньше, меняем только n_clusters.

```
# Ваш код здесь
inertia = []
for k in tqdm_notebook(range(1, n_classes + 1)):
    #
    #
```

Вопрос 5:
Какое количество кластеров оптимально выбрать, согласно методу локтя?

Ответ:

- 1
- 2
- 3
- 4

Попробуем еще один метод кластеризации, который описывался в статье – агломеративную кластеризацию.

```
ag = AgglomerativeClustering(n_clusters=n_classes,
                             linkage='ward').fit(X_pca)
```

Посчитайте Adjusted Rand Index (sklearn.metrics) для получившегося разбиения на кластеры и для KMeans с параметрами из задания к 4 вопросу.

```
# Ваш код здесь
```

Вопрос 6:
Отметьте все верные утверждения.

Варианты:

- Согласно ARI, KMeans справился с кластеризацией хуже, чем Agglomerative Clustering
- Для ARI не имеет значения какие именно метки присвоены кластерам, имеет значение только разбиение объектов на кластеры
- В случае случайного разбиения на кластеры ARI будет близок к нулю

Можно заметить, что задача не очень хорошо решается именно как задача кластеризации, если выделять несколько кластеров (> 2). Давайте теперь решим задачу классификации, вспомнив, что данные у нас размечены.

Для классификации используйте метод опорных векторов – класс sklearn.svm.LinearSVC.

Настройте для LinearSVC гиперпараметр C с помощью GridSearchCV.

- Обучите новый StandardScaler на обучающей выборке (со всеми исходными признаками), прмиените масштабирование к тестовой выборке
- В GridSearchCV укажите cv=3.

```
# Ваш код здесь
#
X_train_scaled =
X_test_scaled =
svc = LinearSVC(random_state=RANDOM_STATE)
svc_params = {'C': [0.001, 0.01, 0.1, 1, 10]}
# Ваш код здесь
best_svc =
# Ваш код здесь
```

Вопрос 7
Какое значение гиперпараметра C было выбрано лучшим по итогам кросс-валидации?

Ответ:

- 0.001
- 0.01
- 0.1
- 1
- 10

```
y_predicted = best_svc.predict(X_test_scaled)
tab = pd.crosstab(y_test, y_predicted, margins=True)
tab.index = ['ходьба', 'подъем вверх по лестнице', 'спуск по лестнице',
             'сидение', 'стояние', 'лежание', 'все']
tab.columns = tab.index
tab
```

Вопрос 8:
Какой вид активности SVM определяет хуже всего в терминах точности? Полноты?

Ответ:

- по точности – подъем вверх по лестнице, по полноте – лежание

- по точности – лежание, по полноте – сидение
- по точности – ходьба, по полноте – ходьба
- по точности – стояние, по полноте – сидение

Наконец, проделайте то же самое, что в 7 вопросе, только добавив PCA.

- Используйте выборки X_train_scaled и X_test_scaled
- Обучите тот же PCA, что раньше, на отмасшабированной обучающей выборке, примените преобразование к тестовой
- Настройте гиперпараметр C на кросс-валидации по обучающей выборке с PCA-преобразованием. Вы заметите, насколько это проходит быстрее, чем раньше.

Вопрос 9:
Какова разность между лучшим качеством (долей верных ответов) на кросс-валидации в случае всех 561 исходных признаков и во втором случае, когда применялся метод главных компонент? Округлите до целых процентов.

Варианты:

- Качество одинаковое
- 2%
- 4%
- 10%
- 20%

Вопрос 10:
Выберите все верные утверждения:

Варианты:

- Метод главных компонент в данном случае позволил уменьшить время обучения модели, при этом качество (доля верных ответов на кросс-валидации) очень пострадало, более чем на 10%
- PCA можно использовать для визуализации данных, однако для этой задачи есть и лучше подходящие методы, например, tSNE. Зато PCA имеет меньшую вычислительную сложность
- PCA строит линейные комбинации исходных признаков, и в некоторых задачах они могут плохо интерпретироваться человеком

Зачетно-экзаменационные материалы для промежуточной аттестации (экзамен)

1. Основные понятия Big Data.
2. Основные методики анализа Big Data.
3. Процесс аналитики анализа Big Data.
4. Особенности хранения Big Data.
5. Характеристика Big Data в мире и России.
6. Определение понятия Data Mining.
7. Определение понятия KDD.
8. Программные средства анализа больших данных.

9. Первичный анализ данных с помощью инструментов визуализации. Очистка данных.
10. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
11. Метрические методы классификации. Метрики качества алгоритмов классификации и регрессии.
12. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.
13. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.
14. Деревья решений. Методы построения деревьев.
15. Случайный лес, его особенности.
16. Градиентный бустинг, его особенности при использовании деревьев в качестве базовых алгоритмов.
17. Нейронные сети. Метод обратного распространения ошибок.
18. Обучение без учителя. Методы понижения размерности. PCA.
19. Обучение без учителя. Кластеризация.
20. Вопросы безопасности Big Data

Критерии оценивания результатов обучения

Оценка	*Критерии оценивания по экзамену*
Высокий уровень «5» (отлично)	*оценку «отлично» заслуживает студент, освоивший знания, умения, компетенции и теоретический материал без пробелов; выполнивший все задания, предусмотренные учебным планом на высоком качественном уровне; практические навыки профессионального применения освоенных знаний сформированы.*
Средний уровень «4» (хорошо)	*оценку «хорошо» заслуживает студент, практически полностью освоивший знания, умения, компетенции и теоретический материал, учебные задания не оценены максимальным числом баллов, в основном сформировал практические навыки.*
Пороговый уровень «3» (удовлетворительно)	*оценку «удовлетворительно» заслуживает студент, частично с пробелами освоивший знания, умения, компетенции и теоретический материал, многие учебные задания либо не выполнил, либо они оценены числом баллов близким к минимальному, некоторые практические навыки не сформированы.*
Минимальный уровень «2» (неудовлетворительно)	*оценку «неудовлетворительно» заслуживает студент, не освоивший знания, умения, компетенции и теоретический материал, учебные задания не выполнил, практические навыки не сформированы.*

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление

информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

5. Перечень учебной литературы, информационных ресурсов и технологий

5.1. Учебная литература

1. Анализ данных : учебник для академического бакалавриата / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2018. — 490 с. — (Бакалавр. Академический курс). — ISBN 978-5-534-00616-2. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/412967

2. Парфенов, Ю. П. Постреляционные хранилища данных : учебное пособие для вузов / Ю. П. Парфенов ; под научной редакцией Н. В. Папуловской. — Москва : Издательство Юрайт, 2023. — 121 с. — (Высшее образование). — ISBN 978-5-534-09837-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/514724 (дата обращения: 09.07.2023).

5.2. Периодическая литература

Указываются печатные периодические издания из «Перечня печатных периодических изданий, хранящихся в фонде Научной библиотеки КубГУ» https://www.kubsu.ru/ru/node/15554, *и/или электронные периодические издания, с указанием адреса сайта электронной версии журнала, из баз данных, доступ к которым имеет КубГУ:*

1. Базы данных компании «Ист Вью» http://dlib.eastview.com
2. Электронная библиотека GREBENNIKON.RU https://grebennikon.ru/

5.3. Интернет-ресурсы, в том числе современные профессиональные базы данных и информационные справочные системы

Электронно-библиотечные системы (ЭБС):

1. ЭБС «ЮРАЙТ» https://urait.ru/
2. ЭБС «УНИВЕРСИТЕТСКАЯ БИБЛИОТЕКА ОНЛАЙН» www.biblioclub.ru
3. ЭБС «BOOK.ru» https://www.book.ru
4. ЭБС «ZNANIUM.COM» www.znanium.com
5. ЭБС «ЛАНЬ» https://e.lanbook.com

Профессиональные базы данных:

1. Web of Science (WoS) http://webofscience.com/
2. Scopus http://www.scopus.com/
3. ScienceDirect www.sciencedirect.com
4. Журналы издательства Wiley https://onlinelibrary.wiley.com/
5. Научная электронная библиотека (НЭБ) http://www.elibrary.ru/

6. Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН http://archive.neicon.ru
7. Национальная электронная библиотека (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) https://rusneb.ru/
8. Президентская библиотека им. Б.Н. Ельцина https://www.prlib.ru/
9. Электронная коллекция Оксфордского Российского Фонда https://ebookcentral.proquest.com/lib/kubanstate/home.action
10. Springer Journals https://link.springer.com/
11. Nature Journals https://www.nature.com/siteindex/index.html
12. Springer Nature Protocols and Methods https://experiments.springernature.com/sources/springer-protocols
13. Springer Materials http://materials.springer.com/
14. zbMath https://zbmath.org/
15. Nano Database https://nano.nature.com/
16. Springer eBooks: https://link.springer.com/
17. "Лекториум ТВ" http://www.lektorium.tv/
18. Университетская информационная система РОССИЯ http://uisrussia.msu.ru

Информационные справочные системы:

1. Консультант Плюс - справочная правовая система (доступ по локальной сети с компьютеров библиотеки)

Ресурсы свободного доступа:

1. Американская патентная база данных http://www.uspto.gov/patft/
2. Полные тексты канадских диссертаций http://www.nlc-bnc.ca/thesescanada/
3. КиберЛенинка (http://cyberleninka.ru/);
4. Министерство науки и высшего образования Российской Федерации https://www.minobrnauki.gov.ru/;
5. Федеральный портал "Российское образование" http://www.edu.ru/;
6. Информационная система "Единое окно доступа к образовательным ресурсам" http://window.edu.ru/;
7. Единая коллекция цифровых образовательных ресурсов http://school-collection.edu.ru/ .
8. Федеральный центр информационно-образовательных ресурсов (http://fcior.edu.ru/);
9. Проект Государственного института русского языка имени А.С. Пушкина "Образование на русском" https://pushkininstitute.ru/;
10. Справочно-информационный портал "Русский язык" http://gramota.ru/;
11. Служба тематических толковых словарей http://www.glossary.ru/;
12. Словари и энциклопедии http://dic.academic.ru/;
13. Образовательный портал "Учеба" http://www.ucheba.com/;
14. Законопроект "Об образовании в Российской Федерации". Вопросы и ответы http://xn--273--84d1f.xn--p1ai/voprosy_i_otvety

Собственные электронные образовательные и информационные ресурсы КубГУ:

1. Среда модульного динамического обучения http://moodle.kubsu.ru
2. База учебных планов, учебно-методических комплексов, публикаций и конференций http://mschool.kubsu.ru/
3. Библиотека информационных ресурсов кафедры информационных образовательных технологий http://mschool.kubsu.ru;
4. Электронный архив документов КубГУ http://docspace.kubsu.ru/

5. Электронные образовательные ресурсы кафедры информационных систем и технологий в образовании КубГУ и научно-методического журнала "ШКОЛЬНЫЕ ГОДЫ" http://icdau.kubsu.ru/

6. Методические указания для обучающихся по освоению дисциплины (модуля)

Самостоятельная работа студентов является неотъемлемой частью процесса подготовки. Дисциплину рекомендуется изучать путем систематической проработки лекционного материала, самостоятельной проработки рекомендуемой литературы, руководств и методических указаний к выполнению практических занятий. Цель самостоятельной работы – расширение кругозора и углубление знаний в области финансового инструментария.

Контроль за выполнением самостоятельной работы проводится при изучении каждой темы дисциплины на семинарских занятиях. Это текущий опрос, тестовые задания, контрольная работа.

В часы, отведенные для самостоятельной работы, студенты под руководством преподавателя обязаны выполнять индивидуальные практические задания, полученные на практических занятиях. При выполнении этих заданий необходимо использовать теоретический материал, делать ссылки на соответствующие формулы, проверять выполнимость предпосылок, необходимых для применения того или иного метода.

В освоении дисциплины инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная учебная работа (консультации) – дополнительное разъяснение учебного материала.

Индивидуальные консультации по предмету являются важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или лицом с ограниченными возможностями здоровья.

7. Материально-техническое обеспечение по дисциплине (модулю)

По всем видам учебной деятельности в рамках дисциплины используются аудитории, кабинеты и лаборатории, оснащенные необходимым специализированным и лабораторным оборудованием.

Наименование специальных помещений	Оснащенность специальных помещений	Перечень лицензионного программного обеспечения
Учебные аудитории для проведения занятий лекционного типа	Мебель: учебная мебель Технические средства обучения: экран, проектор, ноутбук	Microsoft Windows 8, 10, Microsoft Office Professional Plus Loginom Community
Учебные аудитории для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации	Мебель: учебная мебель Технические средства обучения: экран, проектор, ноутбук	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Учебные аудитории для проведения лабораторных работ Лаборатория информационных и управляющих систем 201Н Лаборатория экономической информатики 202Н Лаборатория управления в	Мебель: учебная мебель Технические средства обучения: экран, проектор, компьютеры, ноутбуки Оборудование: ПК, Терминальные станции, Усилитель автономный беспроводной Типовой комплект учебного	 Microsoft Windows 8, 10, Microsoft Office Professional Plus Loginom Community Microsoft Windows 8, 10,

технических системах 207Н	оборудования "Теория автоматического управления", Презентации и плакаты Усилитель автономный беспроводной с микрофоном	Microsoft Office Professional Plus Loginom Community
Лаборатория организационно-технологического обеспечения торговой и маркетинговой деятельности 201А	Панель интерактивная, Конференц-система, Микшер-усилитель, Подавитель акустической обратной связи, Настенный громкоговоритель, Радиосистема, Микрофон на гибком держателе, Моноблок HP, Документ-камера, Беспроводная точка доступа, Система видеоотображения, ЖК панель, Сплитер, Мультимедийная трибуна лектор, Система видеоконференцсвязи, Плакаты	Microsoft Windows 8, 10, Microsoft Office Professional Plus 1С: Предприятие 8 Loginom Community
Лаборатория экономики и управления 212Н	Презентации и плакаты, Многофункциональный профессиональный видео детектор банкнот и ценных бумаг, Счетчики банкнот, Инфракрасный детектор банкнот и ценных бумаг, Универсальный детектор банкнот и ценных бумаг, Детектор подлинности банкнот, Ящик денежный, Планшетный импринтер, Усилитель автономный Беспроводной	Microsoft Windows 8, 10, Microsoft Office Professional Plus
Лаборатория безопасности жизнедеятельности 105А	Лабораторные стенды, Типовой комплект учебного оборудования, Стенды-тренажеры, Стенд-планшет, Тренажерный комплекс по применению первичных средств пожаротушения, Комплекс – тренажер по оказанию первой доврачебной помощи, Робот-тренажер, Комплект плакатов, Комплект демонстрационных пособий, Комплект аудиовизуальных пособий	Microsoft Windows 8, 10, Microsoft Office Professional Plus

Для самостоятельной работы обучающихся предусмотрены помещения, укомплектованные специализированной мебелью, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.

Наименование помещений для самостоятельной работы обучающихся	Оснащенность помещений для самостоятельной работы обучающихся	Перечень лицензионного программного обеспечения
Помещение для самостоятельной работы обучающихся (читальный зал Научной библиотеки)	Мебель: учебная мебель Комплект специализированной мебели: компьютерные столы	Microsoft Windows 8, 10, Microsoft Office Professional Plus

	Оборудование: компьютерная техника с подключением к информационно-коммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду образовательной организации, веб-камеры, коммуникационное оборудование, обеспечивающее доступ к сети интернет (проводное соединение и беспроводное соединение по технологии Wi-Fi)	
Помещение для самостоятельной работы обучающихся (ауд.213 А, 218 А)	Мебель: учебная мебель Комплект специализированной мебели: компьютерные столы Оборудование: компьютерная техника с подключением к информационно-коммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду образовательной организации, веб-камеры, коммуникационное оборудование, обеспечивающее доступ к сети интернет (проводное соединение и беспроводное соединение по технологии Wi-Fi)	Microsoft Windows 8, 10, Microsoft Office Professional Plus